Using Generalizability Theory to Assess Interrater Reliability of Contract Proposal Evaluations

Richard A. Kass, Timothy Elig and Karen Mitchell
US Army Research Institute for the Behavioral and Social Sciences

The purpose of this report is to present a technique for estimating interrater reliability in terms of a generalizability coefficient, give an example of this technique from five recent contract proposal evaluations, and present the implications of these data for organizing future contract proposal reviews.

Generalizability Theory

Most investigations of interrater reliability report the product moment correlation between the ratings of the raters. When more than two raters are employed, the product moment correlation may be reported for all possible pairings of raters. There are three general disadvantages with the correlational approach to assess interrater reliability. First, there is a theoretical problem of conceptualizing proposal evaluation scores in terms of the classical notion of true scores. Second, the correlational method does not permit the investigation of different sources of error. Third, when more than two evaluators are involved, pair-wise correlations do not readily allow for estimates of rater reliability based on composite ratings.

Generalizability Theory is an analysis of variance approach to interrater reliability explicated most completely in a book by Cronbach, Gleser, Nanda and Rajaratnam (1972) entitled The Dependability of Behavioral Measurements. Brennan (1977) provides an amplification of the basic principles and procedures.

The first advantage of Generalizability Theory is that it does not rest on the classical notion of true and error scores. Evaluating contract proposals in terms of classical test theory assumes that there is associated with each proposal a true score, and the more (or better) raters employed the better the final observed score will approximate a proposal's true score. In Generalizability Theory, there is no single true score which the evaluators are attempting to approximate. The Generalizability Coefficient (GC) is an index of how well we are measuring (approximating) one particular specified universe out of any number of possible universes of interest.

A universe is a collection of behavioral measurements. A particular set of behavioral measurements in a universe is further defined in terms of the facets or conditions of measurement. With respect to contract proposal evaluations, there are often three facets: raters, criteria and proposals. It will later be shown that the calculation of the GC on the data in this report involves computing a three-factor (facets) completely crossed ANOVA. The "generalizability" (universe of interest) of Generalizability Theory refers to the extent that the facets defining the universe of interest may be fixed or random.

It will be useful to show the relationship between the calculation of the reliability coefficient (Rxx) and the Generalizability Coefficient (GC).

Reliability can be written as:

$$Rxx = \frac{\sigma^{2}(T)}{\sigma^{2}(T) + \sigma^{2}(E)} \qquad (1)$$

Where T and E represent true and error scores, respectively.
If we substitute universe score U for true score, the equation for the generalizability coefficient (GC) is:

$$GC = \frac{\sigma^{2}(U)}{\sigma^{2}(U) + \sigma^{2}(E)} \qquad (2)$$

It can be seen that the relationship among the terms remains the same for reliability and generalizability coefficients. The major difference is that the relative size of the U and E terms in the GC formulation will vary depending on the number of facets defining the universe score and whether these facets are considered fixed or random facets.

It was stated earlier that the second major limitation of the correlational approach to interrater reliability is its inability to distinguish different sources of error. In classical test theory there is one complex error term. In Generalizability Theory error variance may be identified for each facet. Estimation of the sources of error variance is most useful in making decisions concerning the design of future contract proposal evaluations. One can answer the question of how much interrater reliability would be affected by increasing or decreasing the number of raters or number of criteria, or both.

The third limitation of the traditional correlational approach is that it becomes awkward when more than two raters are used in the evaluation. The traditional approach is to report the product moment correlation between all possible pairings of raters. In some cases an average or median correlation may be given as a single index for the interrater reliability. There are problems with this approach. An individual correlation between any pair of raters represents the reliability of the evaluation score, if either rater's score was used as the proposal's final score. In practice, this is never done. Both raters' scores are used to yield a composite score. Consequently, the correlation between individual rater's scores is an underestimate of the reliability of the composite score. Since all correlations between possible pairs of ratings are underestimates, the average or median of these correlations will be an underestimate also. The extent to which the correlation underestimates the reliability of a composite score increases as the number of raters increases. The Generalizability coefficient provides an index of the reliability of the composite rating. In this manner it may be noted that generalizability coefficients are interclass correlations (Ebel, 1951). Generalizability Theory, however, is an expansion of the interclass coefficient approach to allow for more complex experimental designs.

TABLE 1

A Rater by Criterion by Proposal (PxRxC) ANOVA Design

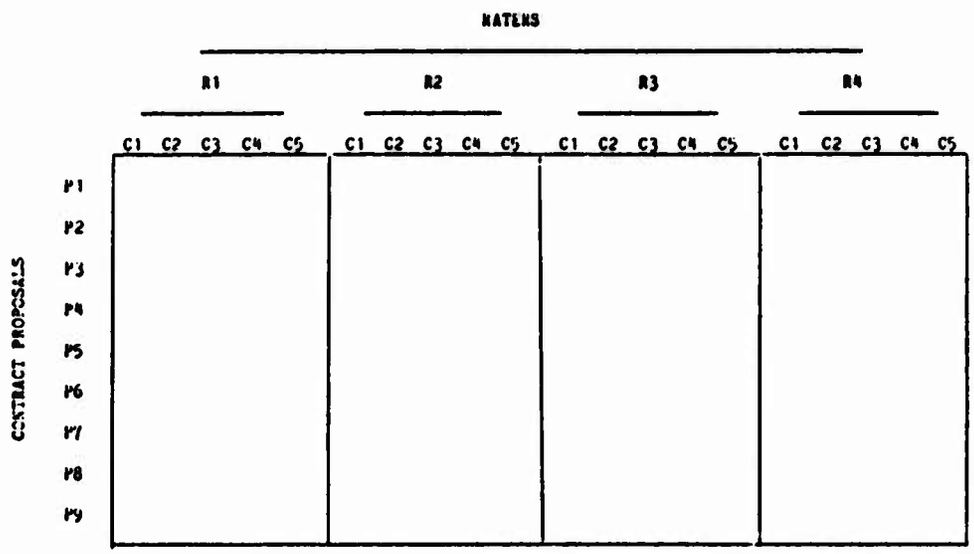

TABLE 2

ANOVA Summary Table for the (PxRxC) Design

Source	df	SS	MS
Proposals (P)	8	791	98.9
Raters (R)**	3	121	40.3
Criterion (C)	4	8,809	2,202.2
PR	24	357	14.8
PC	32	1,015	31.7
RC	12	876	73.0
PRC	96	545	5.7

An Empirical Example

In this section, the interrater reliability of five different sets of contract proposals are analyzed using the generalizability theory approach. The contract evaluations are actual evaluations conducted at the US Army Research Institute (ARI) and they vary along the following dimensions:

Contract Proposal Evaluation Set	Number of Proposals Evaluated	Number of ARI Raters	Number of Criteria Used
A	3	5	3
B	6	3	3
C	8	4	4
D	9	4	5
E	31	3	4

To illustrate the ANOVA method, the interrater reliability of contract proposal evaluations set "D" is worked out in a step-by-step fashion. Table 1 depicts set "D" contract proposals evaluation in terms of a three-way ANOVA experimental design. Nine proposals were received, four raters were used. Each rater (R) rated all proposals (P) with respect to five criteria (C). These criteria reflect separate ratings for different aspects of the proposals, for example, technical adequacy, organizational experience, etc. Accordingly, each proposal received a total of 20 ratings (4 raters x 5 criteria).

In contract proposal evaluations, raters are considered a random facet so that the final evaluation scores will generalize to the use of other raters having similar levels of expertise. The criterion facet is considered a fixed facet in that the final evaluation scores do not generalize to other criteria. That is, the use of some other criteria for a proposal evaluation may result in a different final rank ordering of the proposals.
The proposals facet is considered a random facet in that having more or fewer proposals would not change the score assigned to any one proposal.

Table 2 presents the traditional ANOVA summary data for the actual ratings obtained in the proposal evaluation. In the traditional ANOVA, emphasis is on the statistical tests of the "main" and "interaction" effects by selecting the ratio of the appropriate Mean Square effect and appropriate Mean Square error term. In Generalizability Theory the ANOVA summary table is used only to obtain the quantities for the Mean Squares.

The next step is to compute the unique variance estimates for each facet using data in the ANOVA summary table and the formulations of the components of the Expected Mean Squares. Fortunately, there are well worked out procedures for this (Brennan, 1977). The final variance estimates for the separate facets are presented in Table 3 under the column for G-study variance estimates.

Generalizability theory distinguishes between G studies and D studies. G studies are oriented towards obtaining estimates of the various sources of error variances and G studies are characterized by random-effects ANOVA models. D studies, on the other hand, are designed to determine variance estimates in an

TABLE 3

Changes in Interrater Reliability Due to Changes in the Number or Raters or Criteria

Component	G Study Variance Estimate	Possible D Studies										
		R=4 C=5		R=6 C=5		R=3 C=5		R=4 C=3		R=4 C=7		
		N	σ^2	N	σ^2	N	σ^2	N	σ^2	N	σ^2	
Proposals (P)	2.9	1	2.90	1	2.90	1	2.90	1	2.90	1	2.90	U
Raters (R)	0											
Dimensions (C)	58.3											
PR	1.8	4	.45	6	.30	3	.60	4	.45	4	.45	E
PC	6.5	5	1.30	5	1.30	5	1.30	3	2.17	7	.93	U
RC	7.5											
PRC	5.7	20	.29	30	.19	15	.38	12	.48	28	.20	E
Total Universe Variance (U)			4.20	=	4.20		4.20		5.07		3.83	
Total Error Variance (E)			.74	=	.49		.98		.93		.65	
Generalizability Coeff (GC)			.85	=	.90		.81		.85		.85	

TABLE 4

Computed Generalizability Coefficient
for Each of the Five Contract Proposal Evaluations

Number of Raters

NUMBER OF CRITERIA	3	4	5
3	Data Set B .99		Data Set A .95
4	Data Set E .88	Data Set C .94	
5		Data Set D .85	

actual situation where some facets of the ANOVA model are fixed. While our empirical example is a D study, the results can be used to estimate G-study variances by temporarily assuming that the three facets are random effects. These estimated G-study variances can, in turn, be used to estimate variances for various D-study configurations of interest. The individual D study variance estimates are obtained by dividing the G-study variance estimates by their respective sampling frequencies. The D-study universe (U) and error (E) variances are combined according to equation 2 to compute the GC. For data set "D" with four raters (R = 4) and five criteria (C = 4), the generalizability coefficient is .85.

Extrapolation of Data Set "D" to Other Evaluation Designs

One can compute the extent of expected change in the GC when either (or both) the number of raters or number of criteria is changed. The necessary computation is quite easy. To determine the effect on interrater reliability of increasing the number of raters from four to six, the sampling frequency (N) is changed accordingly and the G-study variances are divided by the new sampling frequencies. This procedure is equivalent to using the Spearman-Brown prophesy formula to determine increases in reliability as test length is increased.

Data in Table 3 summarize changes in the GC for data set "D" when the number of raters or criteria is changed. Increasing or decreasing the number of raters directly increases or decreases the GC. This is because both ANOVA components involving raters contribute to the error term. This may be contrasted to the negligable effect resulting from changes in the number of criteria. Since criteria contribute to both the universe score variance and error variance, the GC ratio of these two terms changes little.

Extrapolation to Other Evaluation Designs Using All Five Data Sets

The projected changes in interrater reliability in Table 3 are based on the G-study variance estimates from one data set. Estimates of the effects of increasing and decreasing the number of raters and/or criteria on interrater reliability are strengthened to the extent that more G-study variance estimates are obtained. The procedure outlined for data set "D" was applied to the other four data sets. The computed generalizability coefficients for all five data sets are presented in Table 4.

The information in Table 4 can be used to compute the effects on reliability of changing the number of raters and/or criteria. Five replications of Table 4 can be estimated by using each data set independently to estimate changes in GC due to changes in the number of raters and criteria. Combining these five sets of independent estimates would yield five interrater reliability coefficients in each cell of Table 4. Moreover, the table can be expanded to provide estimates for combinations of one to seven raters and one to seven criteria. For comparison purposes, the means for each cell have been plotted in Figure 1.

Figure 1 indicates that as the number of raters used in the evaluation increases, so does the interrater reliability. The rate of increase decreases, however, as the number of raters exceeds five. In a similar manner, there is little effect of increasing the number of criteria beyond three. These data suggest for similar evaluations an average level of interrater reliability of .90 can be attained by using three raters and three criteria per contract proposal evaluation.

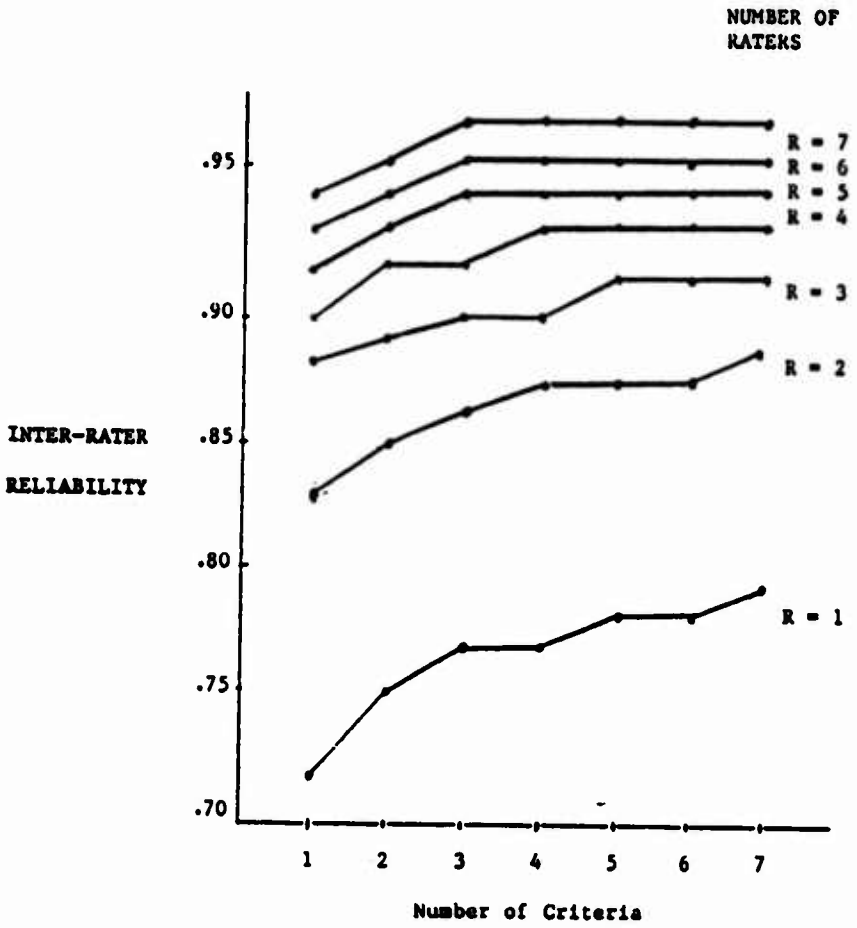

FIGURE 1. Interrater Reliability as a Function of the Number of Raters and Criteria.

REFERENCES

Brennan, R. L. Generalizability analyses: principles and procedures (ACT Technical Bulletin No 26). Iowa City, Iowa: The American College Testing Program, 1977.

Cronbach, L. J. K., Gleser, G. C., Nanda, H.A.N. and Rajaratman, N. The dependability of behavioral measurements. New York: John Wiley and Sons, Inc., 1972.

Ebel, R. L. Estimation of reliability of rating. Psychometrica, 1951, 16, pp. 407-424.